Data 100/200, Spring 2021

Homework 6

Due Date: Thursday, March 11th, at 11:59 PM

Total Points: 40

Submission Instructions

You must submit this assignment to Gradescope by **Thursday, March 11th, at 11:59 PM**. While Gradescope accepts late submissions, you will not receive **any** credit for a late submission if you do not have prior accommodations (e.g. DSP).

You can work on this assignment in any way you like. Here are a few approaches below most students use:

- One way is to use some form of LaTeX. Overleaf is a great tool.
- You could also write your answers on blank sheets of paper. We recommend you use **a new sheet of paper per question** so that it is easier for us to grade your submission.

Regardless of what method you choose, the end result needs to end up on Gradescope, as a PDF. If you wrote something on physical paper (like option 2 above), you will need to use a scanning application (e.g. CamScanner) in order to submit your work.

When submitting on Gradescope, you **must** assign pages to each question correctly (it prompts you to do this after submitting your work). This significantly streamlines the grading process for our tutors. Failure to do this may result in a score of 0 for any questions that you didn't correctly assign pages to. If you have any questions about the submission process, please don't hesitate to ask on Piazza.

Collaborators

Data science is a collaborative activity. While you may talk with others about the homework, we ask that you write your solutions individually. If you do discuss the assignments with others please include their names at the top of your submission.

Properties of Simple Linear Regression

1. (7 points) In Lecture 12, we spent a great deal of time talking about simple linear regression, which you also saw in Data 8. To briefly summarize, the simple linear regression model assumes that given a single observation x, our predicted response for this observation is $\hat{y} = \hat{\theta_0} + \hat{\theta_1}x$.

   In Lecture 12, we saw that the $\hat{\theta_0}$ and $\hat{\theta_1}$ that minimize the average L_2 loss for the simple linear regression model are:

   $$\hat{\theta_0} = \bar{y} - \hat{\theta_1}\bar{x}$$
   $$\hat{\theta_1} = r\frac{\sigma_y}{\sigma_x}$$

   Or, rearranging terms, our predictions $\hat{y}$ are:

   $$\hat{y} = \bar{y} + r\sigma_y\frac{x - \bar{x}}{\sigma_x}$$

   (a) (3 points) As we saw in lecture, a residual e_i is defined to be the difference between a true response y_i and predicted response $\hat{y}_i$. Specifically, $e_i = y_i - \hat{y}_i$. Note that there are n data points, and each data point is denoted by (x_i, y_i).

   Prove, using the equation for $\hat{y}$ above, that $\sum_{i=1}^{n} e_i = 0$.

   (b) (2 points) Using your result from part a, prove that $\bar{y} = \bar{\hat{y}}$.

   (c) (2 points) Prove that $(\bar{x}, \bar{y})$ is on the simple linear regression line.

Geometric Perspective of Least Squares

2. (7 points) In Lecture 13, we viewed both the simple linear regression model and the multiple linear regression model through the lens of linear algebra. The key geometric insight was that if we train a model on some design matrix $\mathbb{X}$ and true response vector $\mathbb{Y}$, our predicted response $\hat{\mathbb{Y}} = \mathbb{X}\hat{\theta}$ is the vector in span($\mathbb{X}$) that is closest to $\mathbb{Y}$.

   In the simple linear regression case, our optimal vector θ is $\hat{\theta} = [\hat{\theta_0}, \hat{\theta_1}]^T$, and our design matrix is

$$\mathbb{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} | & | \\ \mathbb{1} & \vec{x} \\ | & | \end{bmatrix}$$

   This means we can write our predicted response vector as $\hat{\mathbb{Y}} = \mathbb{X}\begin{bmatrix} \hat{\theta_0} \\ \hat{\theta_1} \end{bmatrix}$, and also as $\hat{\mathbb{Y}} = \hat{\theta_0}\mathbb{1} + \hat{\theta_1}\vec{x}$.

   Note, in this problem, $\vec{x}$ refers to the n-length vector $[x_1, x_2, ..., x_n]^T$. In other words, it is a feature, not an observation.

   For this problem, assume we are working with the simple linear regression model, though the properties we establish here hold for any linear regression model that contains an intercept term.

   (a) (3 points) Using the geometric properties from lecture, prove that $\sum_{i=1}^{n} e_i = 0$. *Hint: Recall, we define the residual vector as $e = \mathbb{Y} - \hat{\mathbb{Y}}$, and $e = [e_1, e_2, ..., e_n]^T$.*

   (b) (2 points) Explain why the vector $\vec{x}$ (as defined in the problem) and the residual vector e are orthogonal. *Hint: Two vectors are orthogonal if their dot product is 0.*

   (c) (2 points) Explain why the predicted response vector $\hat{\mathbb{Y}}$ and the residual vector e are orthogonal.

Properties of a Linear Model With No Constant Term

Suppose that we don't include an intercept term in our model. That is, our model is now simply $\hat{y} = \gamma x$, where γ is the single parameter for our model that we need to optimize. (In this equation, x is a scalar, corresponding to a single observation.)

As usual, we are looking to find the value $\hat{\gamma}$ that minimizes the average squared loss ("empirical risk") across our observed data $\{(x_i, y_i)\}, i = 1, \ldots, n$.

$$R(\gamma) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \gamma x_i)^2$$

The normal equations derived in lecture no longer hold. In this problem, we'll derive a solution to this simpler model. We'll see that the least squares estimate of the slope in this model differs from the simple linear regression model, and will also explore whether or not our properties from the previous problem still hold.

3. (4 points) Use calculus to find the minimizing $\hat{\gamma}$. That is, prove that

$$\hat{\gamma} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Note: This is the slope of our regression line, analogous to $\hat{\theta_1}$ from our simple linear regression model.

4. (12 points) For our new simplified model, our design matrix $\mathbb{X}$ is

$$\mathbb{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} | \\ \vec{x} \\ | \end{bmatrix}$$

And so our predicted response vector $\hat{\mathbb{Y}}$ can be expressed as $\hat{\mathbb{Y}} = \hat{\gamma}\vec{x}$. ($\vec{x}$ here is defined the same way it was in Question 2.)

Earlier in this homework, we established several properties that held true for the simple linear regression model that contained an intercept term. For each of the following four properties, state whether or not they still hold true even when there isn't an intercept term. Be sure to explain your answer. You can also provide a counterexample if the statement is false.

(a) (2 points) $\sum_{i=1}^{n} e_i = 0$.

(b) (2 points) The column vector $\vec{x}$ and the residual vector e are orthogonal.

(c) (2 points) The predicted response vector $\hat{\mathbb{Y}}$ and the residual vector e are orthogonal.

(d) (2 points) $(\bar{x}, \bar{y})$ is on the regression line.

(e) (4 points) Complete the coding exercise inside the notebook `hw6_q4e.ipynb` and attach the final plot (or provide a sketch of it) as part of your response. In addition, based on your plot, answer the following 2 questions:

- In your plot, does the simple linear regression model without an intercept term have the same slope as the model with an intercept term?
- Describe one shortcoming for a simple linear regression model without an intercept term.

MSE "Minimizer"

5. (10 points) Recall from calculus that given some function $g(x)$, the x you get from solving $\frac{dg(x)}{dx} = 0$ is called a *critical point* of g – this means it could be a minimizer or a maximizer for g. In this question, we will explore some basic properties and build some intuition on why, for certain loss functions such as the MSE loss, the critical point of the loss will always be the minimizer of the loss.

   Given some linear model $f(x) = \gamma x$ for some real scalar γ, we can write the the mean squared error (MSE) loss of the model f given the observed data $\{x_i, y_i\}, i = 1, \ldots, n$ as

   $$\frac{1}{n}\sum_{i=1}^{n}(y_i - \gamma x_i)^2.$$

   (a) (1 point) Let's break the loss function above into individual terms. Complete the following sentence by filling in the blanks using one of the options in the parenthesis following each of the blanks:

   The MSE loss function can be viewed as a sum of n _____ (linear/quadratic/logarithmic/exponential) terms, each of which can be treated as a function of ____ ($x_i/y_i/\gamma$).

   (b) (3 points) Let's investigate one of the n functions in the summation in the MSE loss function. Define $g_i(\gamma) = \frac{1}{n}(y_i - \gamma x_i)^2$ for $i = 1, \ldots, n$. Recall from calculus that we can use the 2nd derivative of a function to describe its curvature about a certain point (if it is facing concave up, down, or possibly a point of inflection). You can take the following as a fact: A function is convex if and only if the function's 2nd derivative is non-negative on its domain. Based on this property, verify that g_i is a **convex** function.

   (c) (2 points) Briefly explain intuitively in words why given a convex function $g(x)$, the critical points we get by solving $\frac{dg(x)}{dx} = 0$ minimizes g. You can assume that $\frac{dg(x)}{dx}$ is a function of x (and not a constant).

   (d) (3 points) Now that we have shown that each term in the summation of MSE is a convex function, one might wonder if the entire summation is convex given it's a sum of convex functions. While the answer to this for a multivariable function is out of scope for this course, we can still build some intuitions by focusing on single-variable functions.

      i. (2 points) Let's look at the formal definition of convex functions.

      Algebraically speaking, a function $g(x)$ is convex if for any two points $(x_1, g(x_1))$

and $(x_2, g(x_2))$ on the function:

$$g(cx_1 + (1-c)x_2) \leq cg(x_1) + (1-c)g(x_2)$$

for any real constant $0 \leq c \leq 1$.

Intuitively, the above definition says that, given the plot of a convex function $g(x)$, if you connect 2 randomly chosen points on the function, the line segment will always lie on or above $g(x)$ (try this with the graph of $y = x^2$).

Using this definition, show that if $g(x)$ and $h(x)$ are both convex functions, their sum $g(x) + h(x)$ will also be a convex function.

ii. (1 point) Based on what you have shown in the previous part, explain intuitively why the sum of n convex functions is still a convex function when you're adding up more than two convex functions.

(e) (1 point) Finally, explain why in our case that, when we follow the same steps of solving for the critical points to the MSE loss function with respect to the parameter, we are always guaranteed that the solution we find will minimize the MSE loss.